

BLOCK MATCHING PROCESSOR
AND METHOD FOR BLOCK MATCHING MOTION ESTIMATION
IN VIDEO COMPRESSION

5 This application is based on and claims priority from U.S. provisional application
60/218,266 filed July 13, 2000, the contents of which are hereby incorporated by
reference.

BACKGROUND OF THE INVENTION

10

1. Field of the Invention

The present invention relates generally to video compression, and in particular, to a block matching processor for block matching motion estimation.

15

2. Description of the Related Art

Motion estimation which exploits temporal redundancies of an image sequence is a crucial step in video compression. Among diverse motion estimation techniques, a block matching algorithm (BMA) has been adopted in today's popular video coding standards due to computational simplicity with
20 favorable performance. For details of the BMA, refer to D. Gall: 'MPEG: A video compression standard for multimedia algorithm', *Comm. ACM*, 1991, 4, pp. 47-58 [reference 1], P. Pirsch, N. Demassieux and W. Gehrke: 'VLSI architectures for video compression', *Proceedings of the IEEE.*, 1995, 2, pp. 220-246 [reference 2], and V. Bhaskaran and K. Konstantinides: 'Image and Video
25 Compression Standards: Algorithms and Architectures' (Kluwer academic publishers, 1990) 1st edn. [reference 3]. However, the enormous amount of computational requirement of the block matching has been a bottleneck in realizing a compact video encoding system. Hence reducing the amount of motion estimation hardware is the primary issue in designing a cost-effective
30 single chip video encoder.

Regarding the prediction performance, the BMA has a few drawbacks which are mainly caused by employing a fixed block size. A stationary assumption within a block and a simplified translational motion model often
 5 violate real situation in image sequences. These problems could be relieved by using a variable size block matching algorithm (VSBMA). For details of the VSBMA, see M. H. Chan, Y. B. Yu and A. G. Constantinides: 'Variable size block matching motion compensation with applications to video coding', *IEE Proc.*, 1990, 8, pp. 205-212 [reference 4] and F. Defaux and F. Moscheni: 'Motion
 10 estimation techniques for digital TV: A review and a new contribution', *Proceedings of the IEEE*, 1995, 6, pp. 858-876 [reference 5].

In the VSBMA, the choice of the block size has long been shown to be a compromise between several factors. The use of smaller blocks results in higher
 15 adaptivity, but the correlation among blocks cannot be exploited, and thus limits the compression ratio achieved. The user of larger blocks can better exploit the picture correlation as a whole, but the stationary assumption within each block may then be distributed, and the quality as a result suffers. To produce the best result, arbitrary size of block should be used in motion vector predictions
 20 according to the rate-distortion function for video sources. However, considering all the facts of performance, computational efficiency and additional bits for block size information, a small number of block sizes are acceptable in practicable video coding systems.

25 In addition, while the earlier video coding standards such as H.261 and MPEG-1 allow a single mode for the motion vector prediction with a 16×16 macroblock, today's prevalent MPEG-2 which is adopted in worldwide digital TV has more functional choices. For this, see ISO/IEC JTC1/SC29/WG11 and ITV-TS SG 15 EG for ATM video coding: 'MPEG-2 test model 5', Apr. 1993
 30 [reference 6]. According to this document, field prediction mode and special

prediction modes are appended besides frame prediction mode. In field pictures, a 16×16 macroblock is decomposed into two 16×8 blocks, where one corresponds to the odd field and the other to the even field. Special prediction modes refer to $16 \times$ motion-compensation and dual-prime mode are also
 5 concerned on a 16×8 block. Furthermore, in the advanced prediction mode of H.263 and MPEG-4, it is allowed to utilize the motion vectors for 8×8 blocks. See ISO/IEC-JTC1/SC29/WG11 N1908: 'Coding of moving pictures and audio', Oct. 1997 [reference 7] and ITU-T Recommendation H.263: 'Video coding for low bit rate communication', Dec. 1995 [reference 8]. Previous efforts on a block
 10 matching processor have mainly focused on the architecture of fixed block size and single prediction mode. For details, see [reference 2], L. De Vos and M. Stegherr: 'Parameterizable VLSI architectures for the full-search block-matching algorithms', IEEE Trans. on Circuits Syst., 1989, 10, pp. 1309-1306 [reference 9], and S. Chang, J. -H. Hwang and C. -W Jen: 'Scalable array architecture design
 15 for full search block matching', IEEE Trans. on CAS for Video Tech., 1995, 10, pp. 332-343 [reference 10]], D. M. Yang, M. T. Sun and L. Wu: 'A family of VLSI designs for the motion compensation block-matching algorithm', IEEE Trans. on Circuits syst., 1989, 10, pp. 1317-1325 [reference 11], and Y. Jehng and L. Chen and T. Chiueh: 'An efficient and simple VLSI tree architecture motion
 20 estimation algorithms', IEEE Trans. on Signal Processing, 1993, 4, pp. 148-157 [reference 12].

A block matching procedure and a hardware mapping for it on a conventional architecture will be described below. An overall computation flow
 25 in block matching with full-search can be expressed as

$$SAD_{\min} = \text{MAXVALUE}$$

$$V_{\min} = (0,0);$$

$$\text{for } m = -K \text{ to } K-1$$

```

for n = -L to L-1
    SAD(m, n) = 0;
    for i = 0 to M-1
        for j = 0 to M-1
            SAD(m, n) = SAD(m, n) + |x(i, j) - y(i+m, j+n)|;
        endfor
    endfor
    if SAD < SADmin then
        SADmin = SAD(m, n);
        Vmin = (m, n);
    endif
endfor
endfor

```

The widely accepted criterion of block distortion measure is Sum of Absolute Difference (SAD). The operations involved for computing SAD(m, n) and SAD_{min} are associative, and thus the order for exploring the index spaces (I, j) and (m, n) is arbitrary. The block matching computation is massively repetitive and thus suited to be realized in a systolic array processor. See [reference 2], [reference 10] to [reference 12], and S. Y. Kung: 'VLSI array processors' (Eaglewood Cliffs, NJ: Prentice Hall, 1988) 1st edn. [reference 13]. Block matching operations with a systolic array can be expressed as follows.

10 First, in the overall computation flow in block matching with full-search, i and j loops are paralleled and mapped onto hardware. All absolute difference values conforming to one distance measure are calculated concurrently in M×N PEs (Processing Elements).

15 The arrangement of the PE and the computation flow in the systolic array

are illustrated in FIG. 1. FIG. 2 shows an example of conventional two-dimensional systolic array for block matching and the internal structure of the PE (see [reference 2] and [reference 5]). By the conventional architecture, we mean the typical two-dimensional systolic array architecture shown in [reference 2] and [reference 9], which has been the base architecture for a systolic array block matching processor. The PE computes differences between pixels in the current frame X and the previous frame Y and collectively accumulates them to produce the block distortion $SAD(m, n)$ for each matched block whose displacement vector is (m, n) . It is symbolically denoted as AD as shown in FIG. 2 and can be decomposed into two sub-PEs, i.e., A and D, as shown in FIGs. 3B and 3C, after operation shown in FIG. 3A. In FIG. 2, an operator M stands for a comparator shown in FIG. 3D and keeps the minimum block distortion.

FIG. 4 is a detailed block diagram of the PE. One PE 100 includes a difference part 102 with an inverter 108 and an adder 110, an absolute part 104 with an inverter 112 and an exclusive-OR gate 114, and an accumulation part 106 with an adder 116 and a register 118. The accumulation part 106 corresponds to an operator A shown in FIG. 3B, and the difference part 102 and the absolute part 104 correspond to an operator D shown in FIG. 3C. In FIG. 4, pixel data is 8 bits. Reference data representing the pixels of a reference frame and current data representing the pixels of a current frame correspond to X and Y, respectively in Fig. 3A and an intermediate result received from a previous PE corresponds to a in FIG. 3A.

As stated above, the PE AD in systolic mesh is decomposed into individual elements A and D, so that both of them can operate simultaneously to speed up the computation (see [reference 9] and [reference 12]).

To deal with various sizes of blocks at miscellaneous motion vector modes, however, additional special hardware is needed. Therefore, extra area

and control overhead are imposed as constraints.

SUMMARY OF THE INVENTION

5 It is, therefore, an object of the present invention to provide a block matching processor which can flexibly deal with various sizes of matching blocks and miscellaneous motion vector prediction modes of the current video coding standards without extra area and control overhead.

10 The foregoing and other objects can be achieved by providing a block matching processor for flexibly supporting block matching motion estimation at motion vector prediction modes.

15 In the case where matching blocks has sizes being multiples of that of the smallest matching block, a plurality of D-unit arrays generate the absolute values of each smallest size matching block. Each D-unit array has D-units arranged corresponding to the pixels of each smallest size matching block, for calculating the difference between the pixels of a current frame and the pixels of a reference frame, and converting the differences to absolute values. An accumulator is
20 connected to the D-unit arrays, for generating SADs for the smallest size matching blocks and SADs for all the matching blocks of various sizes by tree-like hierarchical addition of the absolute values of the smallest size matching blocks received from the D-unit arrays.

25 In the case where the size of matching blocks used is one of 8×8 , 16×8 and 16×16 pixel sizes, a plurality of D-unit arrays generate the absolute values of 8×8 matching blocks. Each D-unit array has D-units arranged corresponding to the pixels of each 8×8 matching block, for calculating the difference between the pixels of a current frame and the pixels of a reference frame, and converting the

differences to absolute values. An accumulator is connected to the D-unit arrays, for generating SADs (Sum of Absolute Difference) for the 8×8 matching blocks and SADs for a 16×8 matching block by tree-like hierarchical addition of the absolute values of the 8×8 matching blocks received from the D-unit arrays.

5

A method for flexibly supporting block matching motion estimation at motion vector prediction modes for a plurality of matching blocks of pixels having non-uniform sizes that are multiples of a smallest matching block of said plurality of matching blocks, comprises the steps of:

- 10 (a) generating an absolute value of each smallest size matching block in each D-unit array of a plurality of difference unit (D-unit) arrays, wherein said each D-unit array comprising a plurality D-units being arranged to correspond with an arrangement of pixels of said each smallest size matching block in each D-unit array;
- 15 (b) calculating differences between the pixels of a current frame and the pixels of a reference frame by each D-unit, and
- (c) converting the differences calculated in step (b) to absolute values; and
- (d) generating a SAD (Sum of Absolute Difference) for said each
- 20 smallest size matching block and a SAD for all of the plurality of matching blocks of pixels having non-uniform sizes by hierarchically adding the absolute values of the smallest size matching blocks received from the D-unit arrays by an accumulator connected to the D-unit arrays for generating a SAD (Sum of Absolute Difference) for said each smallest size matching block, and
- 25 (e) generating a SAD for all of the plurality of matching blocks of pixels having non-uniform sizes hierarchically adding the absolute values of the smallest size matching blocks received from the D-unit arrays.

In another aspect of the present invention, A method for flexibly

supporting block matching motion estimation at motion vector prediction modes, where a size of matching blocks used is one of 8×8 , 16×8 and 16×16 pixel sizes, said method comprising the steps of :

- (a) generating absolute values of each 8×8 matching block for each D-unit array of a plurality of difference unit (D-unit) arrays having D-units arranged corresponding to the pixels of said each 8×8 matching block;
- (b) calculating a difference between the pixels of a current frame and the pixels of a reference frame,
- (c) converting the differences calculated in step (b) to absolute values;
- (d) generating a SAD (Sum of Absolute Difference) for said each 8×8 matching block and a SAD for a 16×8 matching block by hierarchically adding the absolute values of the 8×8 matching blocks generated in step (a) by the plurality of D-unit arrays from an accumulator connected to the D-unit arrays.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and other objects, features and advantages of the present invention will become more apparent from the following detailed description when taken in conjunction with the accompanying drawings in which:

FIG. 1 illustrates a block matching computation flow in a conventional systolic array architecture;

FIG. 2 illustrates a conventional two-dimensional systolic array for block matching;

FIGs. 3A to 3D illustrate the elements shown in FIG. 2;

FIG. 4 is a block diagram of a conventional PE;

FIG. 5 illustrates a matching block organization and PE array mapping according to an embodiment of the present invention;

FIG. 6 is a block diagram of a PE according to the embodiment of the present invention;

FIG. 7 is a block diagram of a block matching processor according to the embodiment of the present invention;

5 FIG. 8 illustrates a matching block organization and mapping to an A-unit according to the embodiment of the present invention;

FIGs. 9A and 9B illustrate a (3, 2) counter and a (4, 2) counter, respectively;

FIG. 10 illustrates the structure of a first level accumulator (ACC1)
10 according to the embodiment of the present invention;

FIG. 11 illustrates the structure of a second level accumulator (ACC1) according to the embodiment of the present invention;

FIG. 12 illustrates carry bit mapping to the A-unit;

FIG. 13A illustrates additional storages for various sizes of blocks;

15 FIG. 13B illustrates an $N \times N$ block composed of sub-blocks B_0 and B_1 ;

FIG. 13C illustrates the structure of an ACS-like comparator unit according to embodiment of the present invention;

FIG. 13D illustrates the structure of a comparator unit according to the embodiment of the present invention; and

20 FIGs. 14A and 14B illustrate standard cell implementation of 128 PE's arrays in the conventional architecture versus the architecture according to the embodiment of the present invention in the aspects of layout and size.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

25

A preferred embodiment of the present invention will be described hereinbelow with reference to the accompanying drawings. In the following description, well-known functions or constructions are not described in detail since they would obscure the invention in unnecessary detail.

30

The following description is made on a block matching processor according to the present invention which flexibly supports block matching motion estimation at motion vector prediction modes each using one size of matching block among 8×8 , 16×8 , and 16×16 pixel sizes with one hardware.

5

FIG. 5 illustrates a matching block organization and PE array mapping. Each pixel of the current block is associated with one of N^2 ($N \times N$: block size) PEs. Once the current matching block data are loaded into PEs, they will be retained until the given search area is fully explored. A standard 16×16 macroblock 200 consists of two 16×8 blocks, and a 16×8 block can be decomposed into two 8×8 blocks. An 8×8 matching block is the smallest basic processing unit in the block matching processor of the present invention and the 8×8 matching blocks are mapped on PE arrays 204 to 210 of a PE array module 202 in FIG. 5. The 8×8 matching blocks are labeled as A_0 , A_1 , A_2 and A_3 ,
 15 respectively in the 16×16 matching block 200. A search area memory module 212 is divided into four memory banks 214 to 220 to provide data concurrently to the PE arrays 204 to 210 which operate in parallel. For efficient memory accesses, a MUX network 222 is inserted between the search area memory module 212 and the PE arrays 204 to 210m which distributes data to proper PE
 20 arrays. A current block memory 224 for storing the pixel data of a 16×16 current frame provides the current data to the PE array 202 through a bus 226.

FIG. 6 illustrates the structure of a PE according to the embodiment of the present invention. The PE is divided into a difference unit (D-unit) 228 and
 25 an accumulation unit (A-unit) 230 according to a stable delay time and efficient utilization of hardware resources in operation data paths. The D-unit 228 includes inverters 232 and 236, an adder 234, an exclusive-OR gate 238, and a register 240. The inverter 232 inverts reference data representing the pixels of a reference frame. The adder 234 adds current data representing the pixels of the

current frame to the inverted reference data. The inverter 236 inverts a carry-out bit generated during the addition in the adder 234 and feeds the inverted carry-out bit \overline{cout} to the A-unit 230. The exclusive-OR gate 238 exclusive-OR gates the outputs of the adder 234 and the inverter 236. The register 240 stores the output of the exclusive-OR gate 238. The A-unit 230 is composed of a first level accumulator (ACC1) 241 and a second level accumulator (ACC2) 242 for accumulating the output of the D-unit 228 at two levels. In the present invention, the output of only one D-unit 228 is not accumulated but a plurality of absolute values generated from D-unit arrays are accumulated together.

10

FIG. 7 is a block diagram of a block matching processor for a 16×16 macroblock in relation with the above-described PE according to the embodiment of the present invention. Referring to FIG. 7, each of four D-unit arrays 0-3 (244 to 250) corresponds to an array of 8×8 D-units 228 shown in FIG. 6. The four D-unit arrays 0-3 (244 to 250) correspond to the PE arrays 0-3 (204 to 210). Therefore, the D-unit arrays 244 to 250 include D-units which are arranged corresponding to the pixels of 8×8 matching blocks, calculate the differences between the pixels of the current frame and the pixels of the reference frame, and convert the differences to absolute values. An accumulator 230 connected to the D-unit arrays 244 to 250 is composed of ACC1s 252 to 258 and 262 to 268 and on pair of ACC2s 260 and 270. The ACC1s 252 and 254 accumulate absolute values generated from the D-unit array 0 (244), the ACC1s 256 and 258 accumulate absolute values generated from the D-unit array 1 (246), the ACC1s 262 and 264 accumulate absolute values generated from the D-unit array 2 (248), and the ACC1s 266 and 268 accumulate absolute values generated from the D-unit array 3 (250). Each of the ACC1s 252 to 258 and 262 to 268 adds the absolute values of a corresponding 8×4 sub-block divided from the 8×8 matching block and generates an SAD for the sub-block. The ACC2 260 generates an SAD for one 8×8 matching block by adding the SADs of two sub-

blocks generated from the ACC1s 252 and 254, generates an SAD for the other 8×8 matching block by adding the SADs of the other two sub-blocks generated from the ACC1s 256 and 258, and generates an SAD for one 16×8 matching block by adding the SADs of the 8×8 matching blocks. In the same manner, the

5 ACC2 270 generates an SAD for one 8×8 matching block by adding the SADs of two sub-blocks generated from the ACC1s 262 and 264, generates an SAD for the other 8×8 matching block by adding the SADs of the other two sub-blocks generated from the ACC1s 266 and 268, and generates an SAD for one 16×8 matching block by adding the SADs of the 8×8 matching blocks.

10

The architecture of the present invention is flexible in the sense that it is capable of managing various sizes of matching blocks and search ranges. For the variable search ranges, the flexibility and scalability can be attained by employing internal buffers and a proper data distribution network with a

15 cascadable PE array configuration. In this case, the structure of an individual PE is not affected inherently. However, it becomes a different story when the size of a matching block changes. Because the maximum value of the SAD depends on the block size, the bit-width of a storage element for the SAD changes according to the size of the block. Moreover, the PE arrays need to be rearranged for the

20 same account. To resolve this problem, the summation process needs to be separated from the pixel difference computation process in the SAD calculations. For additional motivations which follow, a single PE is divided into a D-unit and an A-unit. By dividing the PE into A- and D-units, SAD for the block of various sizes can be obtained easily from the intermediate nodes of an accumulation tree

25 that form an ACC2 as later described. The intermediate result of an error accumulation increases as the matching process goes to the next PE in the systolic array (see FIG. 2). Hence, the bit-width of a storage element also needs to be increased and the regularity of all the PEs in the array can be broken. To maintain the regularity of all PEs, the bit-width of an accumulator should be

fixed to the maximum value of the SAD, or error accumulators should be separated. Since the former approach wastes storage elements, the latter approach is preferable. Cycle time can be shortened by arranging the computational load. The D-unit calculates only differences and the A-unit does
 5 summations. Thus speedup can be acquired at each storage by distributing the three operations (subtraction, absolute value conversion, and accumulation) for block matching operation. Additionally, critical path and timing slack caused by unbalanced delay time among different computation steps can be actually removed by dividing the computation load.

10

Meanwhile, the operations performed in the D-unit 228 shown in FIG. 6 are subtraction and absolute value conversion. A ripple-carry adder (RCA) is employed for the subtraction of two 8-bit positive pixel values. The delay time of an RCA grows in proportion to the bit-width of input operands, and thus the
 15 overall delay in an 8-bit ripple-carry subtractor can be written as $8\Delta_{FA}$ (where Δ_{FA} denotes 1-bit full-adder delay). In the following to compare the speed with that of a different type of adder, a unit gate-delay is denoted as Δ_G (see I. Koren: 'Computer Arithmetic Algorithms' (John Wiley & Sons Inc. 1993) 1st edn. [reference 14], K. Hwang: 'Computer arithmetic, Principles, Architecture and
 20 Design' (John Wiley & Sons Inc., New York, 1979) 1st edn. [reference 15], and N. H. E. Weste and K. Eshraghian, 'Principles of CMOS VLSI Design' (Addison-Wesley Publishing Company, 1993) 2nd edn. [reference 16]). A rough comparison in terms of gate counts and delays between an RCA and a first carry-lookahead adder (CLA) is listed in Table 1.

25

(Table 1)

Adder type	Number of 2-input gates required	Total delay time in terms of Δ_G
Ripple-carry adder	$\approx 7n$	$\approx 2n\Delta_G$

Carry-lookahead adder	$\approx 13n$ ($n=4m$, $m=1, 2, \dots$)	$\approx (n/2+3) \Delta_G$
-----------------------	--	----------------------------

It is assumed that the CLA consists of four-bit unit groups with a separate carry-lookahead in each group. However, carry-lookahead logic over groups is not considered. For $n=8$, an RCA has about a half gate counts with half speed compared to a CLA. By employing an RCA, additional area advantage can be obtained, because it is more regular than a CLA. Furthermore, since the cycle time is not bounded by 8-bit pixel comparison, the area advantage of an RCA has more meaning than the speed advantage of a CLA.

10 The absolute value conversion is just a 2's complement process for a negative number. A 2's complement conversion can be decomposed into two sub-processes: 1's complement (inversion of all bits) and an increment. Taking a 1's complement is a simple and fast operation of passing an inverter, whereas it takes an adder delay time to increment. In the conventional PE structure shown
15 in FIG. 4, an incrementer is implemented within an accumulation stage by feeding the final carry-bit of the subtractor. Likewise, the D-unit 228 takes a 1's complement of the subtraction result and transmits the inverted final carry-bit, \overline{cout} to the A-unit 230.

20 Now, the A-unit 230 will be described. To deal with the various sizes of matching blocks, the A-unit 230 is built in a hierarchical carry-save adder (CSA) tree. Since each of carry and sum is generated independently, the carry-save addition requires more bit-lines and a final merging adder. However, this area penalty is minor compared to the area gains obtained from the regularity and
25 modularity of CSA trees. Furthermore, it takes constant delay time independent of the input operand bit-width because carry need not be propagated to generate the result.

As shown in FIG. 8, a total of 64 pixels in the 8×8 matching block is divided into two parts and each of them is processed in parallel by the ACC1. The ACC1 is constructed using carry-save (4, 2) counters. Although other sophisticated circuit designs of a (4, 2) counter are possible, a (4, 2) counter 272 made up of two (3, 2) counters 272 and 276 shown in FIG. 9B is considered as shown in FIG. 9A (see [reference 14]). In this case, the delay of the (4, 2) counter 272 can be regarded as about twice that of a (3, 2) counter. The advantages of a (4,2) counter over a (3, 2) counter are higher modularity and regularity in constructing the adder tree, especially for the 2^n (n is a positive integer) number of input operands.

One ACC1 includes a total of 15 (4, 2) counters 278 to 306 and features a quad-tree-like configuration as shown in FIG. 10. A total of 32 pixel-difference results of the 8×4 sub-blocks from the D-unit arrays are accumulated through four levels of (4, 2) counters. The depth of the (4, 2) counter tree is determined considering the delay balance between the D-unit and the A-unit. Like in a (3, 2) counter, the delay time of a (4, 2) counter is independent of the input operand bit-width and can be expressed as $2\Delta_{FA}$. Accordingly the propagation delay of a four-level (4, 2) counter is $8\Delta_{FA}$ ($=4 \times 2\Delta_{FA}$), which is matched closely to the total delay time of the D-unit as intended. The data bit-width increases by one as the accumulation progresses to the next (4, 2) counter because the amount of error to be summed is doubled. Since the ACC1 includes carry-save (4, 2) counters, two registers 308 and 310 should be reserved for both sum and carry.

25

As shown in FIG. 11, two stages of the A-unit constitute an ACC2 which forms a binary-tree structure. In the first stage of addition, four inputs from two ACC1s are added to produce the SAD for an 8×8 size matching block. This process is accomplished by generating the SAD for an 8×8 matching block with

two (4, 2) counters 312 and 318 and two carry-propagate adders (CPAs) 314 and 320 and storing the SAD in registers 316 and 322. The second stage of the ACC2 receives the intermediate results from the first stage and produces the SAD for a 16×8 matching block with a single CPA 324. Therefore, the ACC2
5 generates an SAD for one 8×8 matching block by adding the SADs for two sub-blocks generated from a couple of ACC1s, generates an SAD for the other 8×8 matching block by adding the SADs for the other two sub-blocks generated from another couple of ACC1s, and generates an SAD for a 16×8 matching block by adding the SADs of the 8×8 matching blocks. Because SAD for the block of
10 various sizes such as those of an 8×8 matching block and a 16×8 matching block can be obtained easily from the intermediate nodes of the accumulation tree, the various sizes of matching blocks and search ranges can be supported with a single hardware. For a highly pipelined operation, the delay in the ACC@ is also designed to be matched to that of the D-unit and the ACC1. During synthesizing
15 each logic block, special attention is paid to the delay matching among different computation units through intensive timing simulation.

As described above, the (4, 2) counter array that form a quad-tree-like configuration primarily accumulates the SADs for sub-blocks and the ACC2
20 sums the intermediate results, to thereby obtain the SAD for a larger block.

As mentioned in the D-unit description, the result of the subtraction will be 1's complemented if it is negative and then transmitted to the A-unit with an inverted final carry-put bit, i.e. \overline{cout} . The \overline{cout} is to be used as the carry-in of the
25 accumulation adder to complete the 2's complement process as shown in FIG. 6. In the A-unit architecture according to the present invention, all \overline{cout} bits transmitted from the four D-units cannot be reflected because the (4, 2) counter 272 of FIG. 9B has only one carry-in bit for the four input operands. Appending an additional adder logic to exactly evaluate all \overline{cout} 's might break the structural

regularity and the modularity of an accumulator tree. The problem is resolved considering the trade-off between the accuracy and the hardware cost.

As shown in FIG. 12, the four \overline{cout} bits from the four D-units 228 are
 5 compressed into two bits by OR-ing in pairs and then transmitted to the A-unit. One goes to the carry-in of a (4, 2) counter 332 of FIG. 12 and the other to the LSB position of the carry result of the (4, 2) counter 332. The carry and sum generated from the (4, 2) counter 332 are stored in registers 334 and 336, respectively. Accuracy loss occurs only when the two \overline{cout} 's to be OR-ed by
 10 OR-gates 328 and 330 are all 1s, that is, two pixel-difference values are all negative. The probability of the pixel-difference value being negative is 0.5 because it is random and the distribution function can be assumed uniform. Consequently, the probability of losing the accuracy is 0.25 ($=0.5 \times 0.5$) and the expectation value of the accuracy loss per pixel comparison is 0.125 ($=0.25 \times 1/2$)
 15 in the embodiment of the present invention.

The final step in block matching motion estimation is the comparison of the SADs to find the minimum one. To generate an SAD for a large size matching block, the SADs of small size matching blocks must be added up and
 20 stored in ACC_0 to ACC_{n-1} with adders 342 and registers 344 in FIG. 13. If these processes are attributed to a comparator unit, the A-unit need not produce and store the SADs for matching blocks of different sizes. Fortunately, the SAD comparison is carried out in a block-by-block fashion and thus need not be paralleled massively such as block matching operations. Considering all these
 25 observations, a merged add-compare-select (ACS) logic. For an $N \times N$ matching block composed of two sub-blocks as shown in FIG. 13B, the ACC_0 to ACC_{n-1} of FIG. 13A can be merged to a single logic block by employing a comparator 346 shown in FIG. 13C. Therefore, only the SADs for the sub-blocks B_0 and B_1 are produced in each A-unit and then sent to the comparator unit 346. In the

comparator unit 346, an adder 348 adds $SAD(B_0)$ to $SAD(B_1)$ and a register 350 stores $SAD(B_0) + SAD(B_1)$. A comparator 352 compares the $SAD(B_0) + SAD(B_1)$ with the previous SAD stored in a register 356 and a selector 354 selects the smaller of the SADs. The selected SAD is stored in the register 356.

5 By using the comparator unit 346, arithmetic- and storage-elements for the SADs of various matching blocks are much reduced.

FIG. 13D illustrates a comparator unit 358 constructed by adding a selector 360 to the above-described comparator unit 346 according to the

10 embodiment of the present invention. The selector 360 selects one of SADs for 8×8 matching blocks and an SAD for a 16×8 matching block received from an ACC2, and an SAD for an 16×16 matching block added in an adder 348 and then stored in a register 350 according to a motion vector prediction mode and feeds the selected SAD to the comparator 352. The comparator 352 compares the

15 input SAD with the previous SAD and selects the smaller of the two.

The arrays of 128 PEs of the conventional architecture and the architecture of the present invention are implemented for evaluation. Assuming that each of 16 current- and reference-block data can be loaded in parallel, two-

20 dimensional 16×8 PE arrays are constructed. First, the array architectures are designed using the Verilog hardware description language and then register-transfer-level (RTL) simulations are performed for functional verification. Gate-level synthesis, simulation and backend works such as placement, route ad post-layout simulation are all carried out using the COMPASS Design Navigator CAD

25 tool with LG 5V 0.6 μ m 3-metal layer CMOS technology (see LG Semicon: '0.6 μ m 5-Volt High Performance Library - gvsc650/3', LG Semicon Co., Ltd., 1995 [reference 17]). Although a full-custom design style can best exploit the regularity and the area efficiency of an array processor, it requires considerable design efforts. Thus for fast prototyping and architecture evaluation, standard

cell implementation is employed. FIGs. 14A and 14B illustrate the realized standard cell layouts of chips according to the conventional architecture and the architecture of the present invention for comparison in terms of structure and size. The partitioning of the standard cell blocks follows the guidance rule of the implementation technology according to the target operating frequency. The normal operating clock frequency is selected as 54MHz which is common in typical MPEG-2 video encoders.

To completely check the function and timing of the realized VLSI architecture, the post-layout simulations have been run in all the best, typical and worst case conditions as listed in Table 2. After passing all the tests, the longest operational path delay is obtained in the worst case simulation.

(Table 2)

	Simulation condition		
	Best case	Typical case	Worst case
Ambient temperature	0°C	25°C	70°C
Voltage	5.5V	5V	4.5V
Process	best	typical	worst

15

Table 3 summarizes the architecture evaluation results in terms of area and speed (the longest operational path delay). Due to careful delay matching efforts in the PE design and efficient error accumulation network employing carry-save (4, 2) counter, the PE array of the present invention has 34% less area with 42% speed-up compared to the conventional one.

(Table 3)

	Conventional PE array	Proposed PE array
Array configuration	16×8	16×8

Transistor counts	302856	214021
Area	$240.9 \times 158.6 \text{mm}^2$ (38209mm ²)	$184.8 \times 136.1 \text{mm}^2$ (25157mm ²)
The longest operational path delay (post-layout)	14.22ns	10.08ns

As described above, the novel block matching architecture in 0.6μm CMOS technology is implemented and the functional correctness and timing are verified through intensive simulation along at the entire design levels. VLSI realization shows the efficiency improvement of the architecture of the present invention in both area and speed.

Without additional hardware and control overhead, the architecture of the present invention accommodates matching blocks of various sizes and the miscellaneous motion vector prediction modes of current video coding standards. By decomposing the PE into the D-unit and the two-level A-unit, very high efficiency is achieved in area and speed. The ACS style comparator unit is also devised to achieve further area efficiency.

While the invention has been shown and described with reference to a certain preferred embodiment thereof, it is a mere exemplary application. While 8×8, 16×8 and 16×16 matching block sizes are used in the embodiment of the present invention, the present invention is also applicable to more matching block sizes as far as matching blocks used in motion vector estimation modes are of sizes being a multiple of that of the smallest matching block. Furthermore, the adder 348 of the comparator unit can be incorporated into the ACC2. Finally, a person of ordinary skill in the art understands that equivalent circuit logic may be used which is within the spirit of the invention and scope of the appended claims, and the invention is not limited to the hardware shown in the drawings.

Therefore, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention as defined by the appended claims.